Journal of Biomedicine and Biotechnology • 2005:2 (2005) 65–79 • DOI: 10.1155/JBB.2005.65

RESEARCH ARTICLE

# Early-Stage Folding in Proteins *(In Silico)* Sequence-to-Structure Relation

Michał Brylinski,[1,2] Leszek Konieczny,[3] Patryk Czerwonko,[1] Wiktor Jurkowski,[1,2] and Irena Roterman[1]

[1] *Department of Bioinformatics and Telemedicine, Medical Faculty, Jagiellonian University, Kopernika 17, 31-501 Cracow, Poland*
[2] *Faculty of Chemistry, Jagiellonian University, Ingardena 3, 30-060 Cracow, Poland*
[3] *Institute of Biochemistry, Medical Faculty, Jagiellonian University, Kopernika 7, 31-501 Cracow, Poland*

Received 16 September 2004; revised 3 January 2005; accepted 5 January 2005

A sequence-to-structure library has been created based on the complete PDB database. The tetrapeptide was selected as a unit representing a well-defined structural motif. Seven structural forms were introduced for structure classification. The early-stage folding conformations were used as the objects for structure analysis and classification. The degree of determinability was estimated for the sequence-to-structure and structure-to-sequence relations. Probability calculus and informational entropy were applied for quantitative estimation of the mutual relation between them. The structural motifs representing different forms of loops and bends were found to favor particular sequences in structure-to-sequence analysis.

## INTRODUCTION

Prediction of three-dimensional protein structures remains a major challenge to modern molecular biology. On the one hand, identical pentapeptide sequences exist in completely different tertiary structures in proteins [1]; on the other, different amino acid sequences can adopt approximately the same three-dimensional structure. However, the patterns of sequence conservation can be used for protein structure prediction [2, 3, 4]. Usually, secondary structure definition has been used for ab initio methods as a common starting conformation for protein structure prediction [5]. A large body of experiments and theoretical evidence suggests that local structure is frequently encoded in short segments of protein sequence. A definite relation between the amino acid sequences of a region folded into a supersecondary structure has been found. It was also found that they are independent of the remaining sequence of the molecule [6, 7]. Early studies of local sequence-structure relationships and secondary structure prediction were based on either simple physical principles [8] or statistics [9, 10, 11, 12]. Nearest-neighbor methods use a database of proteins with known three-dimensional structures to predict the conformational states of test protein [13, 14, 15, 16]. Some methods are based on nonlinear algorithms known as neural nets [17, 18, 19] or hidden Markov models [20, 21, 22, 23]. In addition to studies of sequence-to-structure relationships focused on determining the propensity of amino acids for predefined local structures [24, 25, 26, 27], others involve determining patterns of sequence-to-structure correlations [21, 22, 28, 29, 30]. The evolutionary information contained in multiple sequence alignments has been widely used for secondary structure prediction [31, 32, 33, 34, 35, 36, 37, 38]. Prediction of the percentage composition of $\alpha$-helix, $\beta$-strand, and irregular structure based on the percentage of amino acid composition, without regard to sequence, permits proteins to be assigned to groups, as all $\alpha$, all $\beta$, and mixed $\alpha/\beta$ [5, 39].

Structure representation is simplified in many models. Side chains are limited to one representative virtual atom; virtual $C\alpha - C\alpha$ bonds are often introduced to decrease the number of atoms present in the peptide bond [40, 41]. The search for structure representation in other than the $\phi$, $\psi$ angles conformational space has been continuing [42].

Other models are based on limitation of the conformational space. One of them divided the Ramachandran map into four low-energy basins [43, 44]. In another study, all sterically allowed conformations for short polyalanine chains were enumerated using discrete bins

Correspondence and reprint requests to Irena Roterman, Department of Bioinformatics and Telemedicine, Medical Faculty, Jagiellonian University, Kopernika 17, 31-501, Poland, E-mail: myroterm@cyf-kr.edu.pl

called mesostates [45]. The need to limit the confomational space was also asserted [46, 47].

The model introduced in this paper is based on limitation of the conformational space to the particular part of the Ramachandran map. The structures created according to this limited conformational subspace are assumed to represent early-stage structural forms of protein folding *in silico*.

In this paper, in contrast to commonly used base of final native structures of proteins, the early-stage folding conformation of the polypeptide chain is the criterion for structure classification.

Two approaches are the basis for the early-stage folding model presented in this paper.

(1) The geometry of the polypeptide chain can be expressed using parameters other than $\phi$, $\psi$ angles. These new parameters are the $V$-angle—dihedral angle between two sequential peptide bond planes—and the $R$-radius, radius of curvature, found to be dependent on the $V$-angle in the form of a second-degree polynomial. Details on the background of the geometric model based on the $V$, $R$ [48, 49] are recapitulated briefly in "appendix A."

(2) The structures satisfying the $V$-to-$R$ relation appeared to distinguish the part of the Ramachandran map (the complete conformational space) delivering the limited conformational subspace (ellipse path on the Ramachandran map). It was shown that the amount of information carried by the amino acid is significantly lower than the amount of information needed to predict $\phi$, $\psi$ angles (point on Ramachandran map). These two amounts of information can be balanced after introducing the conformational subspace limited to the conformational subspace distinguished by the simplified model presented above. Details on the background of the information-theory-based model [50] are reviewed briefly in "appendix B."

The conformational subspace found to satisfy the geometric characteristics (polypeptide limited to the chain peptide bond planes with side chains ignored) and the condition of information balancing appeared to select the part of Ramachandran map which can be treated as the early-stage conformational subspace.

The introduced model of early-stage folding was extended to make it applicable to the creation of starting structural forms of proteins for an energy-minimization procedure oriented to protein structure prediction. The characteristics and possible applicability of the sequence-to-structure and structure-to-sequence contingency tables is the aim of this paper.

The structures created according to the limited conformational subspace can be reached in two different ways: (1) as the partial unfolding (Figures 1a–1e) and (2) as the basis for the initial structure assumed to represent early-stage folding (Figures 1f–1j). The partial unfolding of the native structural form (called the "step-back" structure in this paper) is expressed by changing the $\phi$, $\psi$ angles to the $\phi_{sb}$, $\psi_{sb}$ angles ($\phi_{sb}$, $\psi_{sb}$ angles belong to the ellipse path, and their values are obtained according to the criterion of the shortest distance between $\phi$, $\psi$ and the ellipse—shown in Figure 1b). The second approach, in which the structure is created on the basis of the $\phi_{es}$, $\psi_{es}$ angles ($\phi_{es}$, $\psi_{es}$ denote the dihedral angles belonging to the ellipse and representing a particular probability maximum), is based on the library of sequence-to-structure relations for tetrapeptides.

A scheme summarizing the two procedures—partial unfolding and partial folding—is shown below (Figure 1). The procedure called partial unfolding starts at the native structure of the protein (Figure 1a). The values of the $\phi$, $\psi$ angles present in the protein are changed (according to the shortest distance criterion) to the values of the angles belonging to the ellipse ($\phi_{sb}$, $\psi_{sb}$). When these dihedral angles are applied, the structure of the same protein looks as is shown in Figure 1c. When this procedure is applied to all proteins present in the protein data bank, a probability profile can be obtained which represents the distribution of $\phi$, $\psi$ angles in the limited conformational subspace. The distribution is different for each amino acid, although some characteristic maxima can be distinguished. The profile shown in Figure 1d represents Glu (the ellipse equation $t$-parameter = 0° represents the point of $\phi = 90°$ and $\psi = -90°$, and then increases clockwise). Particular probability maxima can be recognized using the letter codes also shown in Figure 2. These letter codes are used to classify the structures of proteins in their early-stage folding (*in silico*) (Figure 1e).

The opposite procedure, aimed at protein folding, is shown also in Figures 1f–1j. The starting point in this procedure is the amino acid sequence of a particular protein. After selecting four-amino-acid fragments (in an overlapping system), four different structural codes (for the same tetrapeptide) can be attributed on the basis of the contingency table described above (Figure 1f). Only a particular fragment of the probability profile (according to the letter code) can be recognized in this case. In consequence, the $\phi_{es}$, $\psi_{es}$ values representing the location of the probability maximum on the $t$-axis can be attributed to a particular sequence (Figure 1g). This is why the $\phi_{es}$, $\psi_{es}$ angles differ versus $\phi_{sb}$, $\psi_{sb}$. In consequence, the structure of the transforming growth factor $\beta$ binding protein-like domain (protein selected as an example, PDB ID: 1APJ) created according to the $\phi_{es}$, $\psi_{es}$ angles shown in Figure 1h differs versus the ($\phi_{sb}$, $\psi_{sb}$)-based structure. The "sb" (step-back) and "es" (early-stage) structures differ due to the continuous form of the probability distribution in "sb" procedure and the discrete one in the "es" procedure. The next step in the prediction procedure is energy minimization, which in some cases causes approach toward the native structure (Figure 1j).

The structures created according to the ellipse path treated as the starting structures for the energy-minimization procedure, deliver forms that approach the native structure after one simple optimization procedure. BPTI [51], ribonuclease [50], to some extent also human hemoglobin $\alpha$ and $\beta$ chains [52] and lysozyme [53] were used as the model molecules. All these examples

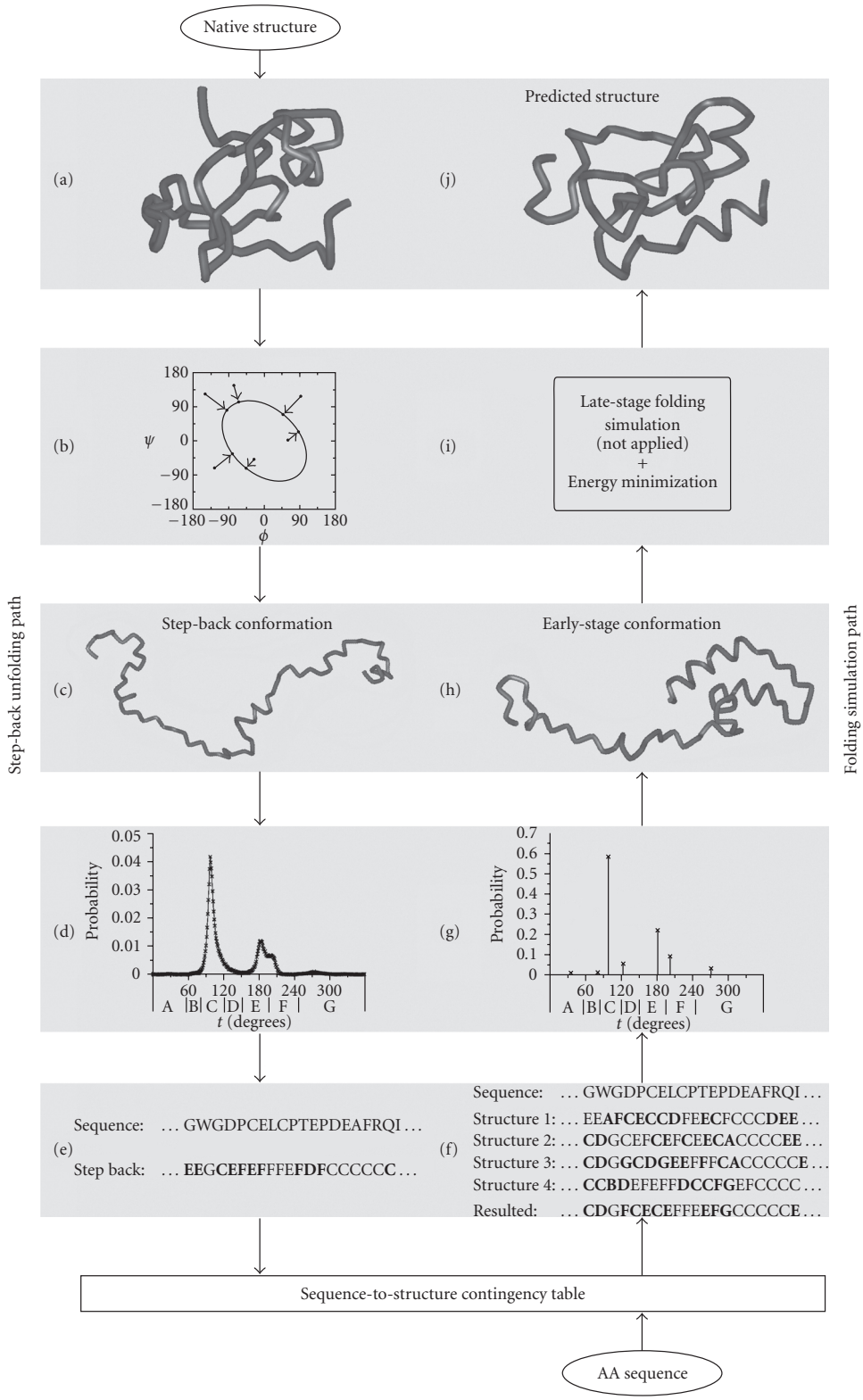



FIGURE 1. (a–e) Step-back unfolding path: (a) native structure of 1APJ, (b) partial unfolding procedure, (c) step-back conformation according to the limited conformational subspace, (d) example of amino-acid-dependent probability profile (Glu) for complete PDB 2003 after moving $\phi$, $\psi$ angles to the nearest point on the ellipse path, (e) letter codes assigned according to probability profiles. (f–j) Folding simulation path: (f) early-stage structure prediction in terms of structural letters, (g) an example of a discrete profile (Glu) applied to early-stage structure creation, (h) predicted early-stage conformation of 1APJ, (i) late-stage folding simulation procedure (under consideration—not applied yet), (j) structure of 1APJ as a result of the energy-minimization procedure with proper disulphide bridges constraints.

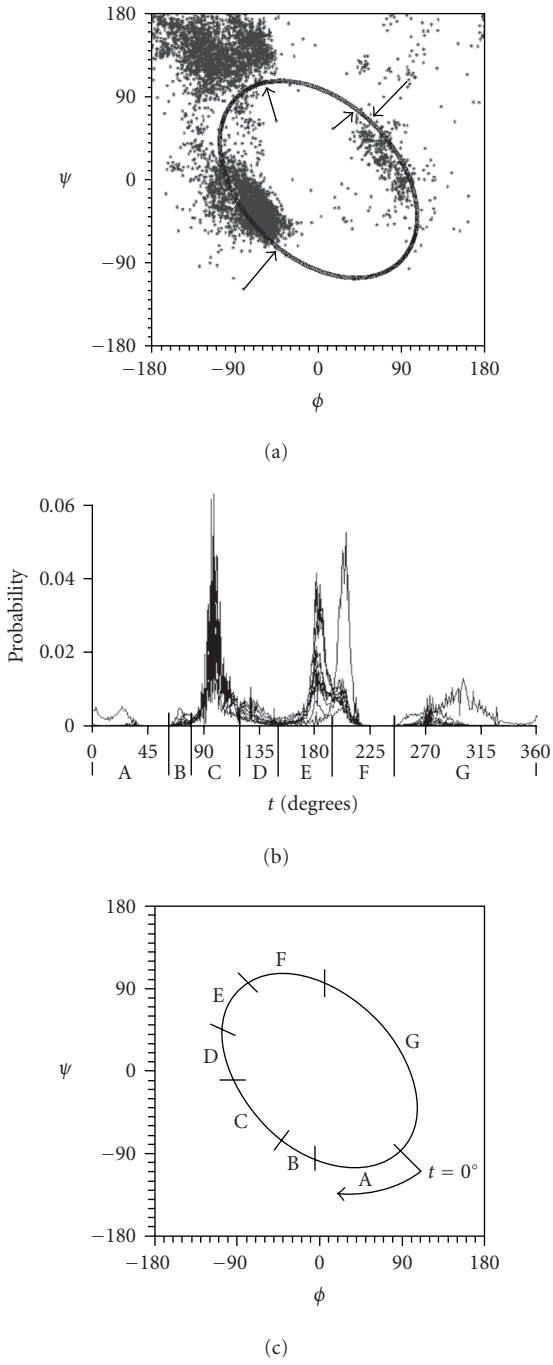



FIGURE 2. Letter codes for structure classification. (a) The ellipse-path-limited conformational subspace in relation to $\phi$, $\psi$ angles as they appear in real proteins. Arrows denote the shortest-distance criterion for definition of $\phi$, $\psi$ angles belonging to the ellipse for arbitrary selected points. (b) Probability maxima as they appear along the ellipse (starting $t$-point shown in (c)) and corresponding letter codes for structure identification. (c) Limited conformational subspace with fragments distinguished according to probability maxima shown in (b).

proved that the ellipse-path-limited conformational subspace helped define the initial structure for the energy-minimization procedure, leading to proper, native-like structures without any forms inconsistent with protein-like ones. When the energy-minimization procedure is not sufficient to deliver the proper native-like structure of the protein (which can be seen in Figures 1a and 1j), the additional procedure is necessary (Figure 1i). It is under study now and will be published in the close future.

## MATERIAL AND METHODS

### *Early-stage folding structure classification*

All proteins present in PDB (release January 2003) were taken for analysis [54]. Letter codes have been used for sequence identification. A letter code system is introduced in this paper for structure representation in protein early-stage folding (*in silico*) based on the probability distribution of $\phi$, $\psi$ angles along the ellipse-path-limited conformational subspace (see "appendix B"). To easily distinguish the structure codes versus sequence codes, the former are printed in bold and the latter in italics in this work.

Comparison of distributions between three-state secondary structures indicated four-amino-acid fragments as the most common ones for $\alpha$-helices, $\beta$-strands, and loops [21, 55]. The tetrapeptide was adopted as the unit for investigation of the sequence-structure relation.

The probability distribution along the ellipse, which is assumed to represent the limited conformational subspace, is the basis for the structure classification introduced in this paper. The profile of the probability distribution (of all amino acids) along the ellipse path is shown in Figure 2. Figure 2a shows the usual distribution of $\phi$, $\psi$ angles as found in proteins together with the ellipse path.

The procedure of moving particular $\phi$, $\psi$ angles to the ellipse path is also shown in Figure 2a. The shortest distance between particular $\phi$, $\psi$ angles (point on the Ramachandran map) and a point belonging to the ellipse path located the $\phi_e$, $\psi_e$ ($e$ denotes ellipse belonging) dihedral angles determining the early stage for a particular amino acid of the polypeptide chain. After moving all $\phi$, $\psi$ angles to the ellipse path, the profile of the probability distribution can be obtained, as shown in Figure 2b. The $t$-parameter is the ellipse parameter present in the equation shown in "appendix A." The $t$-parameter equal to zero represents the point $\phi = 90°$ and $\psi = -90°$ on the Ramachandran map and increases clockwise, as is shown in Figure 2c. Seven probability maxima can be distinguished in this profile. Each of them is letter coded.

This coding system was applied to classify the structures of all proteins analyzed. The codes introduced according to the probability distribution shown in Figure 2b are interpreted as follows: *C* ($t$-value range) represents right-handed helical structures, *E* represents $\beta$-structural forms, and *G* represents left-handed helices. The $\beta$-structural forms are differentiated (some amino acids like Ala, Ser, Asp reveal two probability maxima [50]); this is why code *F* also represents $\beta$-like structures.

Although all other letters represent structural forms not identified in the traditional classification, the presence of probability maxima suggests the need to distinguish these categories (code A mostly for Pro and Gly, code B represented mostly by Asn and Asp, and code D characteristic for Tyr and Asn, to take a few examples).

### The contingency table

A window size of four amino acids (analogous to the open reading frame in nucleotide identification) with one amino acid step (overlapping system) was applied to code the sequences and structures in proteins. Potentially 160 000 ($20^4$) different sequences for tetrapeptides can occur (columns). Taking seven different structural forms for each amino acid in a tetrapeptide, 2401 ($7^4$) structural forms can be distinguished for a tetrapeptide (rows). These numbers give an idea of the size of the contingency table under consideration. For all cells, probability values of $p^t$, $p^c$, and $p^r$ were calculated as follows:

$$p_{ij}^t = \frac{n_{ij}}{N^t}, \tag{1}$$

$$p_{ij}^c = \frac{n_{ij}}{N_j^c}, \tag{2}$$

$$p_{ij}^r = \frac{n_{ij}}{N_i^r}, \tag{3}$$

where $i$ denotes a particular structure (row), $j$ denotes a particular sequence (column), $n_{ij}$ is the number of polypeptide chains belonging to the $i$th structure and representing the $j$th sequence, $N^t$ is the total number of ORFs, and $N_j^c$ and $N_i^r$ denote the number of ORFs belonging to a particular $i$th structure and $j$th sequence, respectively. The table expressing all probabilities ($p_{ij}^t$, $p_{ij}^c$, and $p_{ij}^r$) is available on request at http://www.bioinformatics.cm-uj.krakow.pl/earlystage/. All values are expressed on a logarithmic scale because of the very low probability values in the cells of the table.

### Information entropy as a measure of sequence-to-structure and structure-to-sequence predictability

High values of probability calculated as above (relative to potential probability values) can disclose highly coupled pairs of structure and sequence. Ranking the probability values can extract the highly determined relations for both sequence-to-structure and structure-to-sequence.

Structural predictability can also be measured using informational entropy calculation. According to Shannon's definition [56], the amount of information can be calculated as follows:

$$I_i = N \log_2 p_i, \tag{4}$$

where $I_i$ expresses the amount of information (in bits) dependent on $p_i$—the probability of event $i$. This definition is very useful for measuring the amount of information carried by a particular simple (elementary) event. In the case of a complex event, for which few solutions are possible, informational entropy can be calculated, expressing the level of uncertainty in predicting the solution. Informational entropy according to Shannon's definition is as follows:

$$\mathrm{SE} = -\sum_{i=1}^{n} p_i \log_2 p_i, \tag{5}$$

where $n$ is the number of possible solutions for a particular event. $N$ denotes the number of possible solutions for the event under consideration (number of elementary events).

SE reaches its maximum value for all $p_i$ equal to each other, that is, each $i$th solution is equally probable for the event under consideration and no solution is preferred. The maximum value depends on the number of possible solutions for the event ($n$).

SE equal to zero (or 1.0) represents the determinate case in which only one solution is possible. The higher the difference between $\overset{\max}{\mathrm{SE}}$ and SE, the higher the degree of determinability in the given case. A high $\overset{\max}{\mathrm{SE}} - \mathrm{SE}$ value means that the case is realized by a few solutions and that some of them occur with higher probability, which can be interpreted as a case with higher determinability (biased event).

SE, $\overset{\max}{\mathrm{SE}}$, and the values of the differences between them can be calculated for all rows $\mathrm{SE}^r$ (structural preferences versus amino acid sequence) and for columns $\mathrm{SE}^c$ (sequence preference for a particular structural form) in the contingency table. $\mathrm{SE}^r$ allowed extraction of structures highly determined by the sequence; $\mathrm{SE}^c$ extracted structures highly attributed to a particular sequence.

The SE calculation performed for each column (particular sequence) in the contingency table was calculated as follows:

$$\mathrm{SE}_j^c = -\sum_{i=1}^{N_j^0} p_{ij}^c \log_2 p_{ij}^c, \tag{6}$$

where $\mathrm{SE}_j^c$ denotes informational entropy for the $j$-column, $i$ denotes a particular row (structure), $N_j^0$ is the number of nonzero cells in the $j$-column, and $p_{ij}^c$ is calculated according to (2).

The value $\mathrm{SE}_j^c$ as calculated according to (6) measures the level of uncertainty in predicting structure for the $j$th sequence. The closer the SE value to zero, the higher the degree of chance in prediction.

$\overset{\max}{\mathrm{SE}}$ expresses quantitatively the level of uncertainty in the most difficult case for making a decision. For the $j$-column (sequence):

$$\overset{\max}{\mathrm{SE}}_j^c = -\sum_{i=1}^{N_j^0} \overset{\max}{p_{ij}^c} \log_2 \overset{\max}{p_{ij}^c}, \tag{7}$$

Table 1. Scheme of the sequence-structure contingency table. Symbols explained in text.

| Structure | Sequence | | | | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | ⋯ | $j$ | ⋯ | 160 000 |
| 1 | $n_{11}, p^t_{11}, p^c_{11}, p^r_{11}$ | $n_{12}, p^t_{12}, p^c_{12}, p^r_{12}$ | ⋯ | $n_{1j}, p^t_{1j}, p^c_{1j}, p^r_{1j}$ | ⋯ | $N^r_1$ |
| 2 | $n_{21}, p^t_{21}, p^c_{21}, p^r_{21}$ | $n_{22}, p^t_{22}, p^c_{22}, p^r_{22}$ | ⋯ | $n_{2j}, p^t_{2j}, p^c_{2j}, p^r_{2j}$ | ⋯ | $N^r_2$ |
| ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ |
| $i$ | $n_{i1}, p^t_{i1}, p^c_{i1}, p^r_{i1}$ | $n_{i2}, p^t_{i2}, p^c_{i2}, p^r_{i2}$ | ⋯ | $n_{ij}, p^t_{ij}, p^c_{ij}, p^r_{ij}$ | ⋯ | $N^r_i$ |
| ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ | ⋯ |
| 2 401 | $N^c_1$ | $N^c_2$ | ⋯ | $N^c_j$ | ⋯ | $N^t$ |

where $\overset{\max}{\mathrm{SE}}{}^c_j$ denotes maximum informational entropy for the $j$-column, $i$ denotes a particular row (structure), $N^0_j$ is the number of nonzero cells in the $j$-column, and $\overset{\max}{p}{}^c_{ij}$ denotes the value of probability in a column under the assumption that all nonzero cells are equally represented (the principal condition for $\overset{\max}{\mathrm{SE}}$). In other words, for all nonzero cells ($i = 1, \ldots, N^0_j$) in the $j$-column $\overset{\max}{p}{}^c_{ij}$ can be calculated as follows:

$$\overset{\max}{p}{}^c_{ij} = \frac{1}{N^0_j}. \tag{8}$$

Thus the difference between two quantities ((6) and (7)) can be used as the "distance" between the most difficult situation (all solutions equally possible—random solution) and the situation observed in the case under consideration. For the $j$ column

$$\Delta\,\mathrm{SE}^c_j = \overset{\max}{\mathrm{SE}}{}^c_j - \mathrm{SE}^c_j. \tag{9}$$

Analogous calculations for rows (sequences) were performed. For each $i$-row, the value of $\mathrm{SE}^r_i$, $\overset{\max}{\mathrm{SE}}{}^r_i$, and $\Delta\,\mathrm{SE}^r_i$ was calculated.

## RESULTS

### Structures coded according to the introduced system

Structures of all proteins present in the PDB (release January 2003) [56] were analyzed. The $\phi$, $\psi$ angles were calculated for each amino acid. The $\phi_e$, $\psi_e$ angles were calculated according to the shortest distance versus the ellipse. A letter code was assigned for each amino acid according to the ellipse path fragment. Since the tetrapeptide was used as the structural unit, four letters coded one structural unit. The overlapping reading frame system was applied, which means that one amino acid step was applied in structure classification. The maximum combination of seven letter codes for a four-letter string is equal to 2401. This means that 2401 different four-letter strings were expected to be found. It turned out that only 2397 different strings were found in real proteins. Since there are 20 amino acids and four amino acids were taken for the unit, 160 000 different sequences of tetrapeptides were expected; 146 940 different sequences were found in the proteins under consideration.

### Contingency table

Each tetrapeptide found in proteins was described by a four-letter string expressing the sequence and a four-letter string expressing the structure. Each tetrapetide with a known sequence and known structure can be ordered in the form of a table. The rows of the table represent structures and the columns represent sequences. Finally a $2397 \times 146\,940$ table was constructed. To distinguish the structure codes from sequence codes, sequence codes are in bold capital letters and structure codes in italics. The scheme of the contingency table is presented in Table 1. The total number of tetrapeptides in the analyzed database was found to be 1 529 987. Global analysis of the contingency table shows that the maximum number of different structures attributed to the same tetrapeptide is 144. This tetrapeptide appeared to be of the sequence *GSAA*. The maximum number of different sequences was found for $\alpha$-helix (*CCCC*: 90 587) and for $\beta$-structure (*EEEE*: 47 809). Four structures were not found in the library: *ABAB*, *ABBD*, *ABFB*, *DBAB*.

### Information entropy calculation

SE, $\overset{\max}{\mathrm{SE}}$, and the value of the difference between these two quantities ($\Delta$SE) were calculated according to the procedure presented in "material and methods." They can be calculated for columns (sequences) and for rows (structures) separately. The calculation of $\mathrm{SE}^c_j$ for the $j$-column expresses the information entropy related to the structural differentiation of a particular sequence. The calculation of $\mathrm{SE}^r_i$ for the $i$-row in the contingency table expresses the sequential differentiation for a particular structure. $\overset{\max}{\mathrm{SE}}$ according to information entropy characteristics expresses the entropy for the case in which each of all the nonzero cells represents equal probability. For $\mathrm{SE}^c_j = \overset{\max}{\mathrm{SE}}{}^c_j$, all structures for a particular sequence are equally probable. Equal probability for a set of elementary events (different structures) represents the random situation. The bigger the difference $\overset{\max}{\mathrm{SE}} - \mathrm{SE}$, the more deterministic the case. This is why the difference ($\Delta$ SE) between SE and $\overset{\max}{\mathrm{SE}}$ was taken to measure the degree of structure-to-sequence (or vice versa) determination.

The interpretation of Tables 2 and 3 is as follows. The structural predictability for a particular sequence can be

Table 2. Sequence-to-structure relation measured according to the value of the difference ($\Delta SE^c$) between entropy of information ($SE^c$) calculated for the probability values found in the contingency table (particular column) and maximum entropy of information ($\overset{max}{SE^c}$), which (according to the characteristics of entropy of information) is reached for equal probability values in each nonzero cell in a particular column.

| Sequence | Structure | $SE^c$ (bit) | $\overset{max}{SE^c}$ (bit) | $\Delta SE^c$ (bit) |
|---|---|---|---|---|
| **AAAA** | *CCCC* | 2.29 | 6.44 | 4.15 |
| **GDSG** | *GCFG* | 1.57 | 5.49 | 3.92 |
| **AVRR** | *CCCC* | 1.04 | 4.95 | 3.91 |
| **LAAA** | *CCCC* | 1.77 | 5.61 | 3.84 |
| **EAEL** | *CCCC* | 1.37 | 5.21 | 3.83 |
| **LDKA** | *CCCC* | 1.30 | 5.09 | 3.78 |
| **DAAV** | *CCCC* | 0.69 | 4.46 | 3.77 |
| **AKLK** | *CCCC* | 0.76 | 4.52 | 3.77 |
| **DSGG** | *CFGF* | 1.97 | 5.73 | 3.76 |
| **ELAA** | *CCCC* | 1.30 | 5.04 | 3.75 |

Table 3. Structure-to-sequence relation measured according to the value of the difference ($\Delta SE^r$) between entropy of information ($SE^r$) calculated for the probability values found in the contingency table (particular column) and maximum entropy of information ($\overset{max}{SE^r}$), which (according to the characteristics of entropy of information) is reached for equal probability values in each nonzero cell in a particular column.

| Structure | Sequence | $SE^r$ (bit) | $\overset{max}{SE^r}$ (bit) | $\Delta SE^r$ (bit) |
|---|---|---|---|---|
| *GCFG* | **GDSG** | 4.82 | 7.99 | 3.17 |
| *AEED* | **GLRL** | 3.86 | 6.81 | 2.95 |
| *BACE* | **GGAE** | 2.20 | 5.09 | 2.89 |
| *EAEG* | **IGIG** | 4.79 | 7.68 | 2.89 |
| *AEGE* | **GIGH** | 4.74 | 7.63 | 2.89 |
| *BFBE* | **PEPV** | 2.28 | 5.13 | 2.85 |
| *AEGD* | **GNES** | 2.09 | 4.91 | 2.82 |
| *EBCB* | **ELPD** | 3.68 | 6.38 | 2.70 |
| *EBFB* | **FBEP** | 2.57 | 5.17 | 2.60 |
| *AFFP* | **GFRN** | 2.03 | 4.58 | 2.55 |

estimated in the first case, and the predictability of the sequence for a particular structure in the latter case. The results for only the top ten structures and top ten sequences are shown in Tables 2 and 3.

Its highest structural predictability for a particular sequence confirms polyalanine as a highly probable helical structure. Generally, the highly predictable structures for particular sequences are helical forms (Table 2).

The sequence predictability for particular structural forms displayed a quite unexpected regularity. The structures representing irregular structural forms appeared to reveal the strongest entropy decrease versus the random distribution of sequences. This can be seen analyzing the letter codes for the structures (Table 3).

The top ten structures presented in Table 3 are also shown in Figure 3. In summary, one can say that when a particular irregular structural form is expected in a protein, there are preferable sequences to build these irregular motifs; they are shown in Table 3. This seems to be of particular relevance for threading procedures oriented to the production of new proteins not observed in nature.

## DISCUSSION

Particular classes of amino acid relations to particular structural forms in proteins were recently found to solve the problem of structure predictability [57]. All papers concerning this subject linked sequence with structure as it appears in the final native form of the protein. The model introduced in this paper represents an approach to the relation between sequence and structure in the early-stage folding structural form; the bases for the model are presented in detail elsewhere [48, 49, 50], and verified by BPTI [51], ribonuclease [50], hemoglobin [52], and lysozyme [53] folding. The (*in silico*) early-stage structures of these proteins can be found in the corresponding publications.

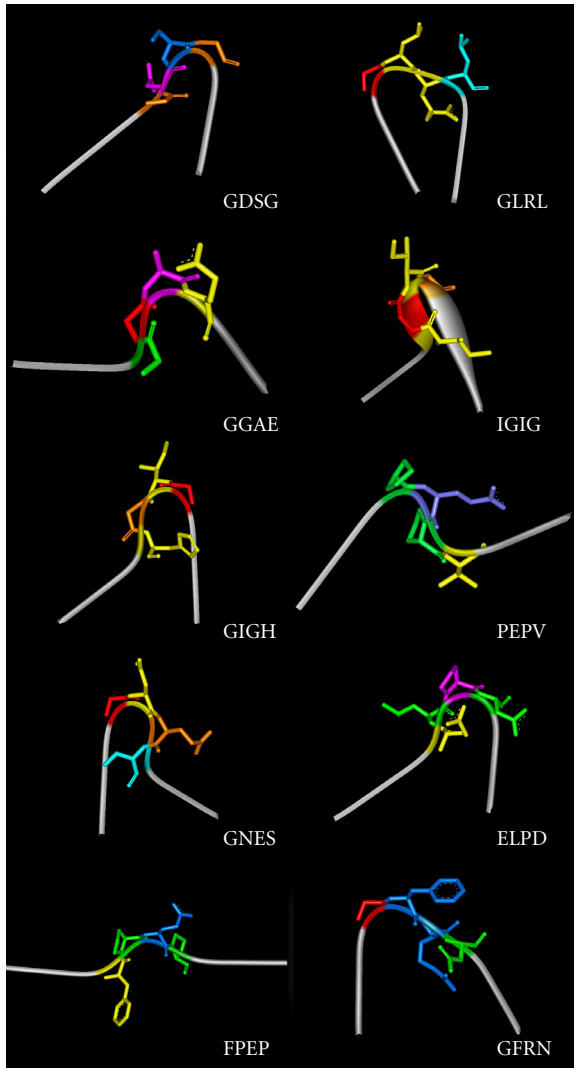



FIGURE 3. Structures of tetrapeptides with highest structure-to-sequence determinability as found using informational entropy calculation (see "material and methods" and Table 1). Gray terminal fragments represent the extended form of polyalanine (tetrapeptides) to emphasize the mutual spatial orientation of terminal fragments. Other colors distinguish ellipse fragments as follows: red (*A*), green (*B*), violet (*C*), sky-blue (*D*), yellow (*E*), dark blue (*F*), orange (*G*). The data for creation of these structures is given in Table 1 and Figure 2.

Several algorithms for quantitatively assigning $\alpha$-helix, $\beta$-strand, and loop regions for proteins with known structure have been developed [58, 59, 60, 61]. The three-dimensional model presented in this paper shows that it is enough to select seven fragments of the ellipse with well-defined probability maxima to be able to predict the early-stage structural form.

The high structure-to-sequence relation found for loops (Table 3, Figure 3) may be particularly important, since a recent survey of 31 genomes indicated that disordered segments longer than 50 residues are very prevalent [62]. Helices, sheets, and turns together account for only about 50%–55% of all protein structure on average [63]; the remaining structures are classified as several types of loops [63, 64]. Current estimations suggest that over 50% of proteins in eukaryotes may carry unconstructed regions of more than 40 residues in length [65], while less than 1% of the proteins in the PDB contains such long disordered regions. These observations taken together imply that many proteins with disordered regions would be unlikely to form crystals [66]. Proteins containing long, disordered segments under physiological conditions are frequently involved in regulatory functions [67], and the structural disorder may be relieved upon binding of the protein to its target molecule [68, 69]. Intrinsically unconstructed proteins and regions, which are also known as natively unfolded and intrinsically disordered, differ from structured globular proteins and domains with regard to many attributes, including amino acid composition, sequence complexity, hydrophobicity, charge, flexibility, and type and rate of amino acid substitutions over evolutionary time [66]. Compared to highly ordered secondary structure regions, the loops and turns are more difficult to identify due to the absence of hydrogen bonding and repeating backbone dihedral angle patterns [70]. The first computational tool indicating the predictability of disordered regions from protein sequence [71] was a neural network predictor (PONDR). Several other disorder predictors have been published since then [72, 73, 74]. Statistically based turn propensity used over a four-residue window was described [75]. The inverse folding problem is the design of protein sequences that have a desired structure [76, 77]. It is impossible to mention even a small part of the papers dealing with the sequence-to-structure relation. Recently, it was concluded that the probability of any state $(\phi,\psi)$ is influenced by the full sequence and not only by the local structure [78].

A genome-scale fold recognition program exploring the knowledge-based structure-derived score function for a particular residue was proposed incorporating three terms: backbone torsion, buried surface, and contact energy [79].

Unlike many others, our model, dual in nature, incorporating sequential and structural information, predicts sequence-to-structure as well as structure-to-sequence.

The contingency table was independently analyzed using another statistics-related method (Meus J, Stefaniak J. The Z coefficient as a measure of dependence in contingency tables (unpublished data), Meus J, Brylinski M, Piwowar P, et al. A tabular approach to the sequence-to-structure relation in proteins (unpublished data)). High accordance was found between the results presented in this paper and in the statistical analysis: the top ten sequences and structures presented in Table 1 were found to be among the most highly correlated, both in sequence-to-structure and in structure-to-sequence, on the ranking list created by the alternate calculation method. The order of the two ranking lists is very similar, additionally confirming the reliability of the model presented.

Aside from early-stage structure prediction, the contingency table presented may contribute to conventional secondary structure prediction, local and supersecondary structure prediction, location of transmembrane regions in proteins, location of genes, or sequence design.

The list of highly determinable tetrapeptides (in sequence-to-structure and structure-to-sequence relations) also allowed the SPI (structure predictability index) scale to be defined [80]. Applied to amino acid sequences, this scale helps to measure the degree of difficulty of structure prediction for a particular amino acid sequence without knowledge of the final, native structure of the protein.

The sequence-to-structure and structure-to-sequence contingency tables, which is created on the basis of all proteins of known structure (step-back procedure), can be used to create the early-stage folding (*in silico*) structure. Applied to other (late-stage folding) procedures, it presumably can enable protein structure prediction. The early-stage form was used as the object for comparison to simplify the presentation of the structure (seven possibilities). The SPI (structure predictability index) parameter, attributed to any amino acid sequence, allows estimation of the degree of difficulty in structure prediction. The probability values (which can be higher or lower) taken from particular cells of the contingency table can tell how offen a particular structure occurs in the protein database so far. The information entropy-based classification presented in this paper allows highly distributed structural forms to be distinguished for a particular tetrapeptide sequence.

## APPENDIX A

The main assumption for the model presented below is that all structural forms of polypeptides in proteins can be treated as helical. The $\beta$-structure in this approach is a helix with a very large radius of curvature. The radius of curvature depends on the $V$-angle, which expresses the dihedral angle between two sequential peptide bond planes. The quantitative analysis of the relation between these two parameters ($V$ and $R$) used the following procedure.

(1) The structure of the alanine pentapeptide was created for each 5° grid point on the Ramachandran map. Each alanine present in the pentapetide represented the $\phi$, $\psi$ angles appropriate for a particular grid point.

(2) Before the parameters ($R, V$) were calculated, all structures (for each grid point) were oriented in a unified way: the averaged position of the carbonyl oxygen atoms and the averaged position of carbonyl carbon atoms determined the $Z$-axis.

(3) The radius of curvature was calculated for projections of $C\alpha$ atoms on the $xy$ plane. The radius of curvature for extended (and $\beta$-structural) forms is very large (theoretically infinite). This is why the natural logarithmic scale was introduced to express the magnitude of $R$.

(4) The $V$-angle was calculated as the difference between the tilt of the central peptide bond plane and the tilt of two (averaged) neighboring peptide bond planes.

The Ramachandram map expressing the $V$-angle distribution and $R$-radius of curvature (in ln scale) is shown in Figure 6.

The ($\ln R$) dependence on the $V$-angle for structures representing low-energy conformations is shown in Figure 4. The approximation function found for this relation is as follows:

$$\ln(R) = 3.4 * 10^{-4} * V^2 - 2.009 * 10^{-2} * V + 0.848. \quad \text{(A.1)}$$

The distribution of $\phi$, $\psi$ angles of structures that satisfy the above equation is shown in Figure 5. The ellipse path found based on this distribution is as follows:

$$\begin{aligned} \phi &= -A\cos(t) - B\sin(t), \\ \psi &= A\cos(t) - B\sin(t), \end{aligned} \quad \text{(A.2)}$$

where $A$ and $B$ are long and short ellipse diagonals, respectively.

## APPENDIX B

The sequence of amino acids in polypeptide determines its structural form. This expression can be understood also as follows. The amount of information carried by an amino acid sequence is comparable to the amount of information necessary to predict its structure.

The amount (bit) of information carried by a particular amino acid can be calculated using Shannon's equation

$$I_i(p_i) = -\log_2 p_i, \quad \text{(B.1)}$$

where $p_i$ expresses the probability of the $i$th amino acid's presence in a sequence.

Assuming all amino acids occur with the same probability (1/20), the amount of information can be calculated.

The amount of information necessary to predict a particular structure (expressed by $\phi_i$, $\psi_i$ dihedral angles) for the $i$th amino acid can also be calculated as follows (using the same Shannon's equation):

$$I_i^{\phi\psi} = -\log_2 p_i^{\phi\psi}, \quad \text{(B.2)}$$

where $p_i^{\phi\psi}$ expresses the probability of the $i$th amino acid to represent the $\phi$, $\psi$ dihedral angles. Assuming 1° as the step for exploring the Ramachandran map and assuming that the Ramachandran map is flat (all $\phi$, $\psi$ angles equally possible), the amount of information $I$ is calculated for $p_i^{\phi\psi}$ equal to $1/(359 * 359)$.

This simple comparison shows that the big difference makes the situation highly nonequilibrated.

The value of $p_i$ is different from 1/20 in real proteins because the frequency of amino acids differs.

The value of $p_i^{\phi\psi}$ also depends on the amino acid under consideration. The assumption of equal probability of

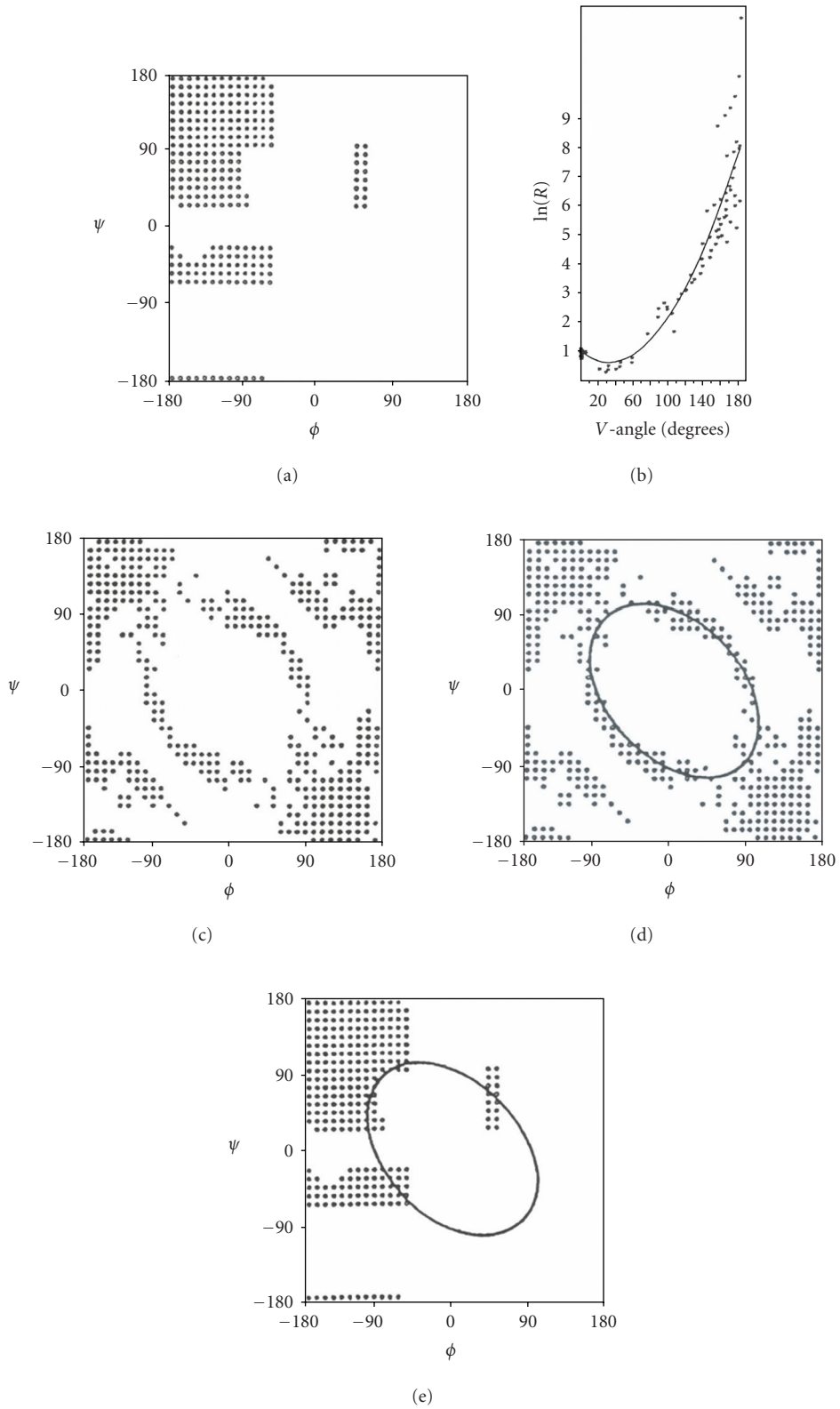
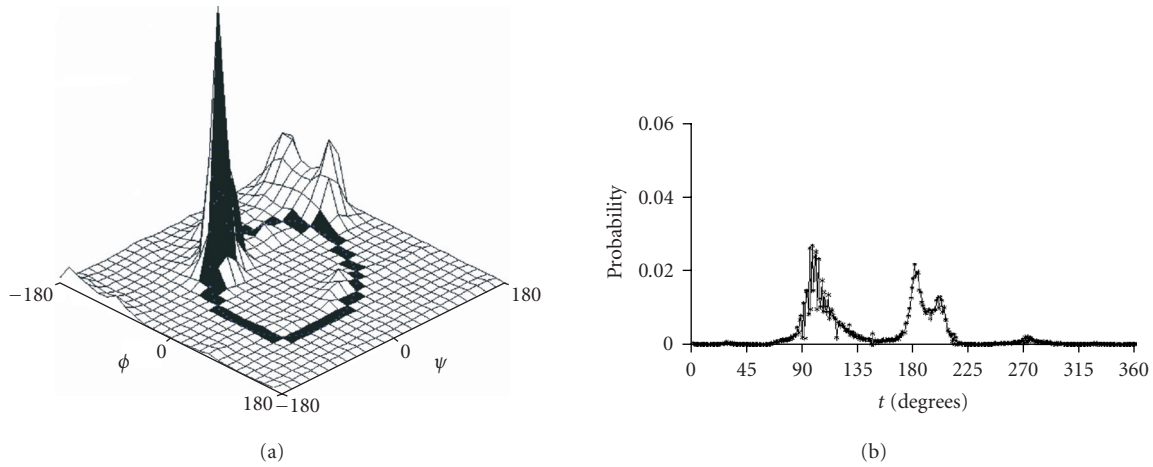
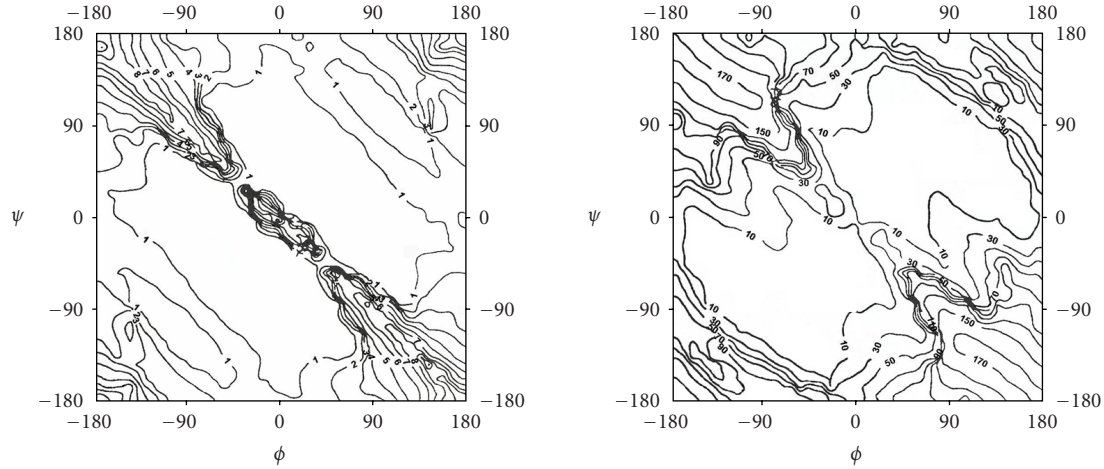



FIGURE 4. Ellipse path determination. (a) $\phi$, $\psi$ map with low-energy area distinguished, (b) $\ln(R)$ as a function of $V$-angle for grid points shown in (a), (c) $\phi$, $\psi$ map with grid points, where the structure satisfies (1), (d) proposed ellipse path, (e) low-energy areas linked by ellipse.




(a) (b)

FIGURE 5. $\phi$, $\psi$ angles distribution of serine: (a) all over the Ramachandran map, black line distinguishes the ellipse path, (b) after moving all $\phi$, $\psi$ angles toward the ellipse path. The variable called $t$ expresses the variable in the ellipse equation (A.2). Zero value of $t$ represents the point $\phi = 90^\circ$, $\psi = -90^\circ$ and then increases clockwise along the ellipse. The probability profiles for each amino acid representing the $\phi$, $\psi$ angles in real proteins after transforming them to the ellipse-path-limited conformational subspace (shortest distance criterion) are presented previously [50].




(a) (b)

FIGURE 6. Distribution of geometrical parameters all over the $\phi$, $\psi$ map. (a) Radius of curvature "$R$" on natural logarithmic scale. (b) Dihedral angle "$V$" between two sequential peptide bond planes.

$\phi$, $\psi$ angles cannot be accepted. Predicting particular $\phi$, $\psi$ angles is relatively easy for proline and most difficult for glycine. Prediction of particular $\phi$, $\psi$ angles is connected with the selection decision. This means selection of $\phi$, $\psi$ from among $359 * 359$ possible solutions. Moreover, particular $\phi$, $\psi$ angles are not equally possible. With information entropy measuring the degree of uncertainty in $\phi$, $\psi$ angles, selection (according to Shannon's equation) is calculated as follows:

$$\mathrm{SE}_i = -\sum_{i=1}^{359*359} p_i \log_2 p_i, \tag{B.3}$$

where index $i$ denotes the amino acid under consideration, $p_i$ denotes the probability of occurrence of particular $\phi$, $\psi$ angles calculated for the $i$th amino acid, $N$ denotes the number of grid points (depending on the step size for $\phi$, $\psi$ angles all over the Ramachandran map), and $\mathrm{SE}_i$ expresses the mean value (quantity) of information (bit) necessary to select one solution from among the number that represents the complete event space ($359 * 359$ in our case). The mean value takes into account the different probabilities for different $\phi$, $\psi$ angles and also the dependence on the amino acid under consideration ($i$th). SE can be interpreted as a scale to measure the predictability

Table 4. Amount of information ($I_i$ (bit)) carried by a particular amino acid, calculated on the basis of the frequency and amount of information ($SE_i^{\phi_e\psi_e}$ (bit), $\phi_e$, $\psi_e$ denote $\phi$, $\psi$ angles belonging to the ellipse) necessary to predict the structure belonging to the ellipse path (early-stage folding conformational subspace) with 10° step of $t$-angle precision (see ellipse equation in "appendix A"). Detailed analysis of the data shown in this table can be found elsewhere [50].

| Amino acid | Amount of information carried by amino acid | Averaged amount of information necessary to predict the ellipse-belonging structure |
|---|---|---|
| | $I_i$ (bit) | $SE_i^{\phi_e\psi_e}$ (bit) |
| Gly | 3.805 | 7.806 |
| Asp | 4.117 | 7.073 |
| Leu | 3.492 | 6.438 |
| Lys | 3.908 | 6.789 |
| Ala | 3.662 | 6.409 |
| Ser | 4.095 | 6.975 |
| Asn | 4.545 | 7.267 |
| Glu | 3.833 | 6.520 |
| Thr | 4.196 | 6.720 |
| Arg | 4.249 | 6.677 |
| Val | 3.886 | 6.233 |
| Gln | 4.663 | 6.676 |
| Ile | 4.151 | 6.208 |
| Phe | 4.713 | 6.617 |
| Tyr | 4.941 | 6.685 |
| Pro | 4.442 | 6.124 |
| His | 5.477 | 6.965 |
| Cys | 5.544 | 6.937 |
| Met | 5.614 | 6.494 |
| Trp | 6.236 | 6.581 |

characteristic for a particular amino acid. It was shown that the SE scale places Gly and Pro at opposite positions on the ranking (scoring) list of amino acids. The $10 * 10$ step for $\phi, \psi$ angles precision prediction still needs a large amount of information to be equilibrated with the amount of information carried by a particular amino acid (in this case $N$ is equal to $35 * 35$).

Analysis of the ellipse path from the point of view of SE calculation reveals that this limited conformational subspace (with 10° steps along the ellipse expressed as $N$ as in (B.4)) satisfies the condition of balancing (Table 4) the amount of information carried by amino acid and the amount of information necessary for selection of the structure belonging to the ellipse path representing the limited conformational subspace with 10° precision.

$$SE_i = -\sum_{i=1}^{360/N} p_i \log_2 p_i, \tag{B.4}$$

where $p_i$ denotes the probability value for a particular point on the ellipse (particular $t$-parameter), and $N$ denotes the number of points selected (it is coupled with the $t$-parameter step size).

The ellipse path presented in "appendix A" appeared to satisfy two important conditions. (i) Almost all structurally important forms of polypeptide are present in this conformational subspace; and (ii) the amount of information carried by the amino acid and the amount of information needed to predict a particular structural form belonging to the conformational subspace are equilibrated. Details on the information problem can be found elsewhere [50]. Figure 5a shows the relation between the $\phi$, $\psi$ angles of Ser distribution all over the Ramachandran map, with the ellipse path distinguished by a black line. The distribution of the $\phi$, $\psi$ angles of Ser after moving them toward the ellipse path is shown in Figure 5b. The overlapping of the probability profiles of all amino acids is shown in Figure 1b.

## ACKNOWLEDGMENTS

We wish to thank Professor Marek Pawlikowski (Faculty of Chemistry, Jagiellonian University) for fruitful discussions. The complete contingency table may be obtained by contacting the authors. This work was financially supported by Collegium Medicum Grants 501/P/133/L, WŁ/222/P/L.

## REFERENCES

[1] Kabsch W, Sander C. On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations. *Proc Natl Acad Sci USA.* 1984;81(4):1075–1078.

[2] Maxfield FR, Scheraga HA. Improvements in the prediction of protein backbone topography by reduction of statistical errors. *Biochemistry.* 1979;18(4):697–704.

[3] Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J Mol Biol.* 1987;195(4):957–961.

[4] Benner SA. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv Enzyme Regul.* 1989;28:219–236.

[5] Shortle D. Prediction of protein structure. *Curr Biol.* 2000;10(2):49–51.

[6] Efimov AV. Role of connections in the formation of protein structures, containing 4-helical segments. *Mol Biol (Mosk).* 1982;16(2):271–281.

[7] Efimov AV. A novel super-secondary structure of proteins and the relation between the structure and the amino acid sequence. *FEBS Lett.* 1984;166(1):33–38.

[8] Lim VI. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol.* 1974;88(4):873–894.

[9] Chou PY, Fasman GD. Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. *Biochemistry.* 1974;13(2):211–222.
[10] Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry.* 1974;13(2):222–245.
[11] Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol.* 1978;120(1):97–120.
[12] Garnier J, Robson, B. The GOR method for predicting secondary structures in proteins. In: Fasman GD, ed. *Prediction of Protein Structure and the Principles of Protein Conformation.* New York, NY: Plenum Press; 1989:417–465.
[13] Biou V, Gibrat JF, Levin JM, Robson B, Garnier J. Secondary structure prediction: combination of three different methods. *Protein Eng.* 1988;2(3):185–191.
[14] Levin JM, Robson B, Garnier J. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.* 1986;205(2):303–308.
[15] Nishikawa K, Ooi T. Amino acid sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods. *Biochim Biophys Acta.* 1986;871(1):45–54.
[16] Yi TM, Lander ES. Protein secondary structure prediction using nearest-neighbor methods. *J Mol Biol.* 1993;232(4):1117–1129.
[17] Holley LH, Karplus M. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA.* 1989;86(1):152–156.
[18] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol.* 1999;292(2):195–202.
[19] Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol.* 1988;202(4):865–884.
[20] Asai K, Hayamizu S, Handa K. Prediction of protein secondary structure by the hidden Markov model. *Comput Appl Biosci.* 1993;9(2):141–146.
[21] Bystroff C, Thorsson V, Baker D. A hidden Markov model for local sequence-structure correlations in proteins. *J Mol Biol.* 2000;301(1):173–190.
[22] Bystroff C, Shao Y. Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. *Bioinformatics.* 2002;18(1):54–61.
[23] Stultz CM, White JV, Smith TF. Structural analysis based on state-space modeling. *Protein Sci.* 1993;2(3):305–314.
[24] Aurora R, Srinivasan R, Rose GD. Rules for alpha-helix termination by glycine. *Science.* 1994;264(5162):1126–1130.
[25] Harper ET, Rose GD. Helix stop signals in proteins and peptides: the capping box. *Biochemistry.* 1993;32(30):7605–7609.
[26] Presnell SR, Cohen BI, Cohen FE. A segment-based approach to protein secondary structure prediction. *Biochemistry.* 1992;31(4):983–993.
[27] Zhou HX, Lyu P, Wemmer DE, Kallenbach NR. Alpha helix capping in synthetic model peptides by reciprocal side chain-main chain interactions: evidence for an N terminal "capping box". *Proteins.* 1994;18(1):1–7.
[28] Bonneau R, Strauss CE, Rohl CA, et al. De novo prediction of three-dimensional structures for major protein families. *J Mol Biol.* 2002;322(1):65–78.
[29] Bystroff C, Baker D. Prediction of local structure in proteins using a library of sequence-structure motifs. *J Mol Biol.* 1998;281(3):565–577.
[30] Han KF, Baker D. Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc Natl Acad Sci USA.* 1996;93(12):5814–5818.
[31] Crawford IP, Niermann T, Kirschner K. Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of tryptophan synthase. *Proteins.* 1987;2(2):118–129.
[32] Russell RB, Breed J, Barton GJ. Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Lett.* 1992;304(1):15–20.
[33] Benner SA, Cohen MA, Gerloff D. Predicted secondary structure for the Src homology 3 domain. *J Mol Biol.* 1993;229(2):295–305.
[34] Levin JM, Pascarella S, Argos P, Garnier J. Quantification of secondary structure prediction improvement using multiple alignments. *Protein Eng.* 1993;6(8):849–854.
[35] Hansen JE, Lund O, Nielsen JO, Brunak S, Hansen JE. Prediction of the secondary structure of HIV-1 gp120. *Proteins.* 1996;25(1):1–11.
[36] Salamov AA, Solovyev VV. Protein secondary structure prediction using local alignments. *J Mol Biol.* 1997;268(1):31–36.
[37] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol.* 1993;232(2):584–599.
[38] Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. *J Mol Biol.* 1994;235(1):13–26.
[39] Aloy P, Stark A, Hadley C, Russell RB. Predictions without templates: new folds, secondary structure, and contacts in CASP5. *Proteins.* 2003;53(6):436–456.
[40] Liwo A, Czaplewski C, Pillardy J, Scheraga, HA. Cummulation-based expression for the multibody terms for the correction between local and electrostatic interaction in the united residue force field. *J Chem Phys.* 2001;115:2323–2347.
[41] Liwo A, Arlukowicz P, Czaplewski C, Oldziej S, Pillardy J, Scheraga HA. A method for optimizing potential-energy functions by a hierarchical design of the potential-energy landscape: application

to the UNRES force field. *Proc Natl Acad Sci USA.* 2002;99(4):1937–1942.
[42] Mezei M. A novel fingerprint for the characterization of protein folds. *Protein Eng.* 2003;16(10):713–715.
[43] Fernandez A, Colubri A, Appignanesi G, Burastero T. Coarse semiempirical solution to the protein folding problem. *Physica A.* 2001;293:358–384.
[44] Sosnick TR, Berry RS, Colubri A, Fernandez A. Distinguishing foldable proteins from nonfolders: when and how do they differ? *Proteins.* 2002;49(1):15–23.
[45] Pappu RV, Srinivasan R, Rose GD. The flory isolated-pair hypothesis is not valid for polypeptide chains: implications for protein folding. *Proc Natl Acad Sci USA.* 2000;97(23):12565–12570.
[46] Colubri A. Prediction of protein structure by simulating coarse-grained folding pathways: a preliminary report. *J Biomol Struct Dyn.* 2004;21(5):625–638.
[47] Alonso DO, Daggett V. Molecular dynamics simulations of hydrophobic collapse of ubiquitin. *Protein Sci.* 1998;7(4):860–874.
[48] Roterman I. Modelling the optimal simulation path in the peptide chain folding—studies based on geometry of alanine heptapeptide. *J Theor Biol.* 1995;177(3):283–288.
[49] Roterman I. The geometrical analysis of peptide backbone structure and its local deformations. *Biochimie.* 1995;77(3):204–216.
[50] Jurkowski W, Brylinski M, Konieczny L, Wiiniowski Z, Roterman I. Conformational subspace in simulation of early-stage protein folding. *Proteins.* 2004;55(1):115–127.
[51] Brylinski M, Jurkowski W, Konieczny L, Roterman I. Limited conformational space for early-stage protein folding simulation. *Bioinformatics.* 2004;20(2):199–205.
[52] Brylinski M, Jurkowski W, Konieczny L, Roterman I. Limitation of conformational space for proteins—early-stage folding simulation of human $\alpha$ and $\beta$ hemoglobin chains. *TASK-Quarterly.* 2004;8:413–422.
[53] Jurkowski W, Brylinski M, Konieczny L, Roterman I. Lysozyme folded in silico according to the limited conformational sub-space. *J Biomol Struct Dyn.* 2004;22(2):149–158.
[54] Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res.* 2000;28(1):235–242.
[55] Zhu ZY, Blundell TL. The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J Mol Biol.* 1996;260(2):261–276.
[56] Shannon CEA. A mathematical theory of communication. *Bell Syst Tech J.* 1948;27:379–423.
[57] Solis AD, Rackovsky S. Optimally informative backbone structural propensities in proteins. *Proteins.* 2002;48(3):463–486.
[58] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 1983;22(12):2577–2637.
[59] Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins.* 1988;3(2):71–84.
[60] Levitt M, Greer J. Automatic identification of secondary structure in globular proteins. *J Mol Biol.* 1977;114(2):181–239.
[61] Sklenar H, Etchebest C, Lavery R. Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins.* 1989;6(1):46–60.
[62] Dunker AK, Lawson JD, Brown CJ, et al. Intrinsically disordered protein. *J Mol Graph Model.* 2001;19(1):26–59.
[63] Leszczynski JF, Rose GD. Loops in globular proteins: a novel category of secondary structure. *Science.* 1986;234(4778):849–855.
[64] Ring CS, Kneller DG, Langridge R, Cohen FE. Taxonomy and conformational analysis of loops in proteins. *J Mol Biol.* 1992;224(3):685–699.
[65] Vucetic S, Brown CJ, Dunker AK, Obradovic Z. Flavors of protein disorder. *Proteins.* 2003;52(4):573–584.
[66] Iakoucheva LM, Dunker AK. Order, disorder, and flexibility: prediction from protein sequence. *Structure (Camb).* 2003;11(11):1316–1317.
[67] Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry.* 2002;41(21):6573–6582.
[68] Dyson HJ, Wright PE. Coupling of folding and binding for unstructured proteins. *Curr Opin Struct Biol.* 2002;12(1):54–60.
[69] Dyson HJ, Wright PE. Intrinsically unconstructed proteins: re-assessing the protein structure-function paradigm. *J Mol Biol.* 1999;293:321–331.
[70] Fetrow JS, Palumbo MJ, Berg G. Patterns, structures, and amino acid frequencies in structural building blocks, a protein secondary structure classification scheme. *Proteins.* 1997;27(2):249–271.
[71] Romero P, Obradovic Z, Kissinger CR, et al. Thousands of proteins likely to have long disordered regions. *Pac Symp Biocomput.* 1998;3:437–448.
[72] Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? *Proteins.* 2000;41(3):415–427.
[73] Liu J, Rost B. NORSp: predictions of long regions without regular secondary structure. *Nucleic Acids Res.* 2003;31(13):3833–3835.
[74] Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 2003;31(13):3701–3708.
[75] Hutchinson EG, Thornton JM. A revised set of potentials for beta-turn formation in proteins. *Protein Science.* 1994;3:2207–2216.
[76] Fischer N, Riechmann L, Winter G. A native-like ar-

tificial protein from antisense DNA. *Protein Eng Des Sel.* 2004;17(1):13–20.

[77] Wie Y, Hecht MH. Enzyme-like proteins from an unselected library of designed amino acid sequences. *Protein Eng Des Sel.* 2004;17:67–75.

[78] Keskin O, Yuret D, Gursoy A, Turkay M, Erman B. Relationships between amino acid sequence and backbone torsion angle preferences. *Proteins.* 2004;55(4):992–998.

[79] Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins.* 2004;55(4):1005–1013.

[80] Brylinski M, Konieczny L, Roterman I. SPI—structure predictability index for proteins. *In Silico Biology.* 2004;5:0022.